

DOCUMENT RESUME

ED 300 431

TM 012 458

AUTHOR Keene, John M.
TITLE Norm-Referenced Test Customization: A Publisher's Viewpoint.
PUB DATE 8 Apr 88
NOTE 23p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 6-8, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Viewpoints (120) -- Speeches/Conference Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Elementary Secondary Education; Equated Scores; Item Banks; Mastery Learning; Minimum Competency Testing; *Norm Referenced Tests; Publishing Industry; Scores; *State Programs; *Test Construction; Testing Programs; Test Interpretation; Test Theory
IDENTIFIERS *Test Customization

ABSTRACT

Increasing numbers of state-mandated testing programs have resulted in the selection of norm-referenced tests (NRTs) for statewide assessment. Adaptation or customization of existing tests has become a viable solution for individual programs. Changes in test instruments have arisen from the behaviorist approach and mastery learning, minimum competency testing, and a cognitive approach. Specific test types arising from each of these concepts are discussed. The national trend has been toward increased involvement of curriculum and instructional personnel in testing. Achievement tests may take varying forms depending on what is to be assessed and the designated outcome. Types of tests available are: (1) shelf NRTs; (2) shelf criterion-referenced tests (CRTs); and (3) customized NRTs or CRTs. Interpretation of test results must always consider that: (1) test equating methods do not mean that interpretation of scores can be the same between a NRT and its customized derivative; (2) content differences occur; (3) individual versus group interpretation varies; and (4) differential instructional effects occur over time. Customized tests provide many benefits but problems of score interpretation must be recognized, as illustrated in a graph. (SLD)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED300431

NORM-REFERENCED TEST CUSTOMIZATION:
A PUBLISHERS VIEWPOINT

John M. Keene

The Psychological Corporation

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as
received from the person or organization
originating it.
- ☐ Minor changes have been made to improve
reproduction quality.
- Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

JANE CANNER

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

A Paper Presented at the
Annual Meeting of the National Council on Measurement in Education
in New Orleans, Louisiana, April 8, 1988

TM012458

Introduction

In recent years assessment of achievement has become increasingly important to schools and communities. As educational practice and focus is pushed further into the limelight, assessment procedures have been put under increased scrutiny. Twenty years ago a school district would select a standardized test and use the results to examine pupil and district success. Almost all assessment was directed and initiated by the school districts.

In the past two decades state legislatures have become increasingly interested in local educational policies. In the assessment area this has resulted in state mandated testing programs. In many states this has resulted in the selection of norm-referenced tests to be used in state-wide assessment programs. In other states assessment purposes have been more narrowly defined and assessment instruments have been developed from state specifications.

State testing programs have created a need for large-scale assessment with quality instruments that reflect state specifications. The adaptation of existing norm-referenced tests has become a viable solution to state and large district needs. This adaptation or customization can range from changing word spellings for a Canadian customization to the complete replacement of items and the development of derivative norms. With the large variety of possibilities customized testing can cover a range of solutions to assessment problems.

This paper examines the background of the recent increase in test customization and provides a discussion of the issues involved in the interpretation of customized test scores.

Changes in Assessment Approaches

Increasingly, standardized achievement tests are being used for policy-oriented purposes: assessing educational equity; providing evidence on school and program effectiveness; allocating compensatory funds to school districts; evaluating teacher effectiveness; accrediting school districts; classifying students for remediation; and certifying successful completion of high school or a given grade of elementary school (Airasian and Madaus, 1983). As achievement testing has assumed increasing importance in effecting educational policy, and as test results have received greater attention from the public, there have been (1) changes in the nature of the tests, and (2) shifts in the types of personnel who are making and enforcing decisions about testing.

There have been philosophical and policy shifts in curriculum and instruction in terms of approach, emphasis, and description which have dictated changes in test instruments. Three major areas of development are identified:

1. The Behaviorist Approach and Mastery Learning

As descriptions of assessment goals have been more refined, the goals of instruction have tended more and more to be expressed in similar terms. Behavioral objectives appeared, combining a specification of curriculum content, relevant behavior, and an acceptable standard of performance. Popham describes techniques for writing amplified objectives...that precisely constrain instructional goals. Bloom's Taxonomy, and Bloom, Hastings, and Madaus in their Handbook on Formative and Summative Evaluation of Student Learning, helped to reinforce the notion that objective tests could be written to assess virtually any educational objective (Haertel and Calfee, 1983). At the same time, Benjamin Bloom advanced the notion of "mastery learning" which says that all students can learn most of what we have to teach in schools given sufficient opportunity and time to learn. In any event, these approaches all require the partitioning of the curriculum into a collection of precise objectives and include a set of test items to measure each objective. In such an environment, the task for teacher and student is to move through the set of items, with instructional materials being chosen and the pace of instruction being decided by the student's performance in the system.

This type of philosophy or approach has resulted in the creation of alternatives to the traditional norm-referenced test:

- a) Norm-referenced tests with objective-based interpretive scores. In one fashion or another, all commercial test publishers have attempted to have their more recently developed test(s) yield objective- or cluster-based scores which identify a student's performance on a given objective or skill in addition to providing typical normative-type scores on a global (test or subtest) level.
- b) Published criterion-referenced tests (CPTs). These tests measure a predetermined set of objectives, such as with the traditional NRT, but the level of specificity of the objectives is greater than most NRTs so as to be more "diagnostic", and a pupil's performance is compared to a fixed criterion rather than to the performance of a reference group.
- c) Customized norm-referenced and criterion-referenced tests. There are at least three approaches which may be used to "customize" an otherwise pre-existing "shelf" NRT. In all approaches, the intent is to modify the shelf test so as to achieve a higher level of congruence between the curriculum and the objectives of the test while retaining both objective-based and normative type scores. In some cases, a new type of score - a Curriculum Referenced score - may be obtained. As with NRTs, technology exists whereby a "shelf" CRT may be modified so as to achieve a higher level of congruence or match between what is being taught and what the test measures.

d) Custom tests from an item bank. An item bank may be defined as a relatively large collection of easily accessible test questions. "Relatively large" means that the number of items exceeds by several times the number to be used in any one test. "Easily accessible" means that the items are indexed, structured, or otherwise assigned information that can be used to facilitate their selection for a test (Millman and Arter, 1984). Such an item bank can be locally developed, or may be obtained via published (commercial or public) sources. Only objectives-referenced type data is usually obtainable, yet methodology exists whereby norm-referenced tests can be constructed, and CRTs can produce estimated or empirical norms.

2. Minimum Competency Approach

Minimum competency testing for pupil certification was born from a number of perceptions: (1) that the pupils' basic skills of literacy and numeracy had seriously deteriorated; (2) that pupils were automatically being promoted from grade to grade regardless of achievement; (3) that pupils did not have the necessary skills to survive in society; and (4) that the high school diploma had become devalued and meaningless as a credential. These perceptions grew out of reports of declining test scores and complaints from businesses, colleges, and the military that disturbing numbers of high school graduates lacked basic skills. Underlying these perceptions was the implicit assumption that most pupils should and can master these skills (Airasian and Madaus, 1983). Hence, an increased emphasis was placed upon building tests that would assess these basic skills. There appears to have been at least three responses in testing that have evolved:

a) Tests measuring "core" or "essential" skills in major content areas. Educators set about the task of identifying those skills and objectives within each discipline that represented the most basic level of expectation, and tests were created (commercial and public) that measured these "core" skills. A more recent emphasis has been the resurgence of interest in secondary level course/subject tests. Instead of focusing just on minimal skills, the process emphasizes the identification of the range of objectives and intended outcomes associated with a given course or subject matter area.

b) Tests measuring writing. It appears that a renewed emphasis on the teaching of writing has occurred during this period of the nation's focus on minimum competencies. Questions about pupil literacy extended to the realm of writing. Commercial and public sources have created tests whereby students write in response to a "prompt" or "rubric." The results are scored "holistically" (judgement of overall impression or intent or quality), or "analytically" (judgement relative to specific criteria), or both.

3. Cognitive Approach

The view of complex learning now extant emphasizes the organization, representation and use of knowledge over the long haul, not just its short-term accumulation...School achievement is no longer to be understood as simply the accretion of facts and content-specific skills...(Rather) a significant part of the learner's task is continually to assemble, reassemble, structure, and tune the cumulating body of knowledge in functional systems for use in thought, and in further learning. Thus, achievement is as much an organization function as it is an acquisition function (Snow, 1980, pp42-43). Hence, educators are placing a renewed emphasis on the development of a coherent structure of knowledge (a "schema") and on procedures for applying and transferring that knowledge to various situations (Haertel and Calfee, 1983). The impact on testing to date appears to be as follows:

- a) There is a renewed interest in tests of reasoning aptitudes as might be measured by verbal or nonverbal mental ability tests.
- b) Bloom's Taxonomy highlighted the dominance in achievement tests (of the time) of items assessing low-level knowledge to the exclusion of higher-level intellectual responses (Haertel and Calfee, 1983). There is an increased emphasis on the inclusion of "higher-order" skills in more recent norm-referenced tests.
- c) There is emphasis on the development of tests specifically focused on higher-order thinking (HOTS). In general, thinking skills are seen as a problem-solving and reasoning strategies, techniques, and habits of thought that may be applied to a variety of types of problems.

Changes in Control of Test Programs

Through the years from 1920 to 1960, standardized norm-referenced achievement tests grew in general use, and they were the dominant form of test used in districts for evaluating pupil and program performance. For the most part, decisions about tests were made at the local rather than state level, and persons at the local level who had the most training in psychometrics and assessment - guidance counselors, educational psychologists, or in larger districts, a director of research - were given the responsibility of planning and executing testing activities. In many communities this is still the case. However, shifts are occurring. As tests have become increasingly important in the role of determining curriculum and instructional effectiveness, and as curriculum and instruction personnel involved themselves in the issues mastery learning, minimum competencies and cognitive approaches as previously identified, it was a natural consequence that curriculum specialists began to provide more input into test development, selection and use. At both local and state department levels, nationally the trend has been that responsibility for testing has been transferred directly to the Curriculum and Instruction Division, or at least curriculum and instruction personnel have much more impact upon the process of assessment and evaluation within the system.

Additionally, with the increase in public (lay and legislative) awareness of educational performance, the concomitant pressure for progress (toward excellence) has fueled a shift in decision-making about test development, selection and use from the local to the state level. Not only are state departments of education becoming more directly involved in the concerns and practices of testing at the local level, but more and more states are mandating state-wide assessment programs. The nature of these programs involve the full range of mastery learning, minimum competency, and cognitive approaches as previously described in this report.

Alternate Approaches to State Testing

In deciding upon the assessment and evaluation approach that will best serve a state's needs, it is important to address a few straightforward issues and questions. Essentially, achievement tests may take a variety of forms depending upon (1) the nature of what is to be measured and (2) the designated outcomes of measurement.

1. What Is To Be Assessed?

Beyond identification of the domains to be tested, is achievement to be measured by examining more global and transferable skills, or is achievement to be based on the measurement of school-specific skills? Is the curriculum defined by a collection of precise objectives and is the intent to measure with test items that are closely or identically matched to the content-specific curriculum, or is the curriculum defined and assessed by procedures for applying and transferring knowledge to various situations? Or is there to be a mix of the two approaches?

Central to this same issue is the question of how closely a test's content should "match" the local/state curriculum. On the one hand, we can insure a match, but in doing so risk a reduction in the importance of learning to apply skills and knowledge to new problems. On the other hand, we can emphasize application and problem-solving by presenting new problems on the test, but risk failing to measure what is being taught (Linn, 1983). Furthermore, in studies of differential school or program effectiveness, there are situations in which we may want to use tests that are not geared to the instructional program to assess opportunity costs or the unintended positive and negative outcomes associated with the program (Airasian and Madaus, 1983).

Additionally, should the content measured be the content of which is prescribed in the state's scope and sequence, or based upon the content which is actually taught in the classroom? When tests are used in studies of school or program effectiveness, the concern is with group performance and generally it is sufficient for the test to show content and curricular validity, i.e., the test content is linked to the state/local educational objectives and to the content of the curriculum materials used in the schools. However, when a test is used to certify individual performance, adequate opportunity to learn the material tested may have to be evaluated on a case-by-case basis to be fair to each pupil.

The test must show instructional validity, i.e., the match between test content and the content of the instruction actually provided pupils in classrooms. In cases where the test's outcomes may impose sanctions upon the student (promotion/retention, graduation), it may be necessary to prove that the test is instructionally valid. In the *Debra P. vs. Turlington* (1979, 1981) case in Florida, the Fifth Circuit ruled that the state had to demonstrate "whether the schools are providing students with instruction in the knowledge and skills measured by the test." In short, the ruling means that a test used as a graduation or promotion requirement must cover material actually taught pupils.

2. What Are The Outcomes of Measurement? Why Test?

The purposes for testing provide the foundation upon which a testing program is built. Without a clear understanding of the purposes for testing, the results are likely to receive minimal attention and use. Listed below are a number of typical purposes for testing along with some associated issues or questions.

- a. To diagnose pupil strengths and needs. How specific must objectives and skills be measured and reported to support a diagnostic function? An increase in specificity typically yields greater diagnostic utility, but testing time generally is increased and the demands upon teachers for the management of information are greater.

- b. To select students for special programs. What are the appropriate selection criteria? What types of scores, objective-referenced or normative, might best represent the criteria? How should the discrepancy standard be set for program admittance (i.e., given economic and program opportunities/constraints, how many students can be allowed into the program)?
- c. To certify pupil performance/progress. How is performance to be judged and reported? By comparison to an objective-based criterion (and how is the criterion to be determined) or comparison to a reference group (which reference group)? What are the consequences of attainment or nonattainment of expected goals?
- d. To develop, implement, monitor and evaluate programs and to allocate resources. For tests to adequately represent this function, they typically must exhibit content and curricular validity. Therefore, what content is to be measured? (Refer to the issues identified in the preceding section: "What is to be assessed?") Also, will it be most meaningful to evaluate on the basis of objective-based scores or on the basis of normative comparisons, or both?
- e. To evaluate and certify school performance. If a test is used as an administrative device and if the sanctions or rewards associated with test performance are important, then the test can have enormous influence over what is taught, how it is taught, what is learned and how it is learned. Such tests can shift control over the curriculum and teaching to the agency that controls the test. The tests can be used to promote educational leadership creating positive effects in curriculum and instruction, or they can devalue or ignore aspects of schooling thereby narrowing and restricting the curriculum (Airasian and Madaus, 1983). Therefore, it becomes a critical task to identify what is to be measured, to determine the criteria for judging success/failure, and to judge the consequences (rewards/sanctions) of attainment or nonattainment of standards.

A special mention ought to be made regarding the issue of setting "fixed" school performance standards. Schools vary considerably with respect to what may be termed as "input variables" - the socio-economic level of parents, "turnover" enrollment or transience of pupils, absenteeism due to out-of-school factors, cultural/language differences, et.al.; therefore, holding all schools to one standard where significant between-school differences exist may not be an entirely fair judgement or evaluation. Yet, for credentialing purposes or particularly when it comes to reporting outcomes to various publics/audiences, a fixed standard is generally required or expected. Therefore, a state must weigh its priorities, and determine whether fixed or variable standards (or both) are necessary.

- f. To report educational achievement and progress to concerned audiences. Obviously, it comes as no surprise that there have been increased demands for the reporting of educational performance, and educators have fed the public's desire for accountability through testing. However, states must address certain questions. Should achievement be reported only on a state level, or should there be release of information on a district-by-district basis? What type of data is most suitable to report - performance represented by objective-based scores or by normative comparisons (or both); static data or changes-over-time? If normative scores are used, what score types are most accurate and easy-to-interpret?

Certainly, there are more reasons why states and districts test than have been listed so far; however, the scope of the needs presented here are typical and representative. Certain types of tests have resulted that are in use in state programs across the nation.

Solutions: Types of Tests

1. Shelf Norm-Referenced Tests (NRT)

As commercial standardized norm-referenced achievement tests are intended to be administered to pupils across the nation, the content topics and skill areas included in the tests are selected to assess the "common denominator" of available curricular materials and content so as to be fair and representative across most schools and pupils. Given that the test content represents, to a fair degree, the nature of the state's expectations about what should be measured, the test can be used for identifying pupil strengths and weaknesses, curriculum and program evaluation, establishment of standards, and reporting to the public. The test may be administered to all students if pupil data is required or the state may utilize a number of different sampling procedures if only aggregate data is required to meet intended outcomes.

2. Shelf Criterion-Referenced Test (CRT)

Like shelf NRTs, shelf criterion-referenced tests measure a predetermined set of objectives representing the "common denominator" of objectives within the domain. CRTs typically measure objectives that are defined on a more specific basis than as compared to NRTs. Testing time is usually longer viz-a-viz NRTs in any given domain. Generally, only objective-based (criterion-referenced) scores are available, but some tests have been "linked" in one fashion or another to some other norm-referenced test thereby providing estimated or derivative norms. States have selected shelf CRTs when the CRT objectives provide a reasonable match to the state's expectations about what should be measured in a particular domain, and when there is an emphasis on diagnosis of individual pupil strengths and needs. Shelf CRTs may be used for curriculum and program analysis outcomes and for reporting to publics as well.

3. Customization of NRTs or CRTs

Test customization involves the development of items or the use of a pre-existing bank of items, or both, to build a test to specific objectives so that it has a 100% match to local/state curricula and courses of study. Most customized tests are criterion-referenced, but it is possible to equate these tests, to norm-referenced tests to provide norms for the custom test. Custom tests are developed, obviously, when shelf tests do not exist or do not provide an adequate representation of the state's expectations about what should be measured. Also, custom tests promote a certain sense of "ownership" that sometimes is as important to develop as the tests themselves. Customization is the most expensive method for building a test, however, and the investment of personnel in terms of time and attention-to-task may be considerable.

Usually the test customization process is undertaken in order to yield more information from one assessment. The specificity and amount of information would not be available from a shelf norm-referenced assessment. Standardized achievement tests are designed to measure fairly broad, generally accepted goals (Mehrens, 1984). Such tests focus on the curriculum most often taught in classrooms across the nation and therefore do not match all local curricula equally well. Significant segments of a particular state or district curriculum may be assessed only minimally in the standardized test, assessed at another grade, or not assessed at all (content-not-tested). Further, the test may include items measuring objectives which are not included in the local curriculum at the particular grade being tested or at any grade for that matter (content-not-taught). On the other hand, a customized test, designed specifically to assess a particular local curriculum, does not usually have norms against which national comparisons can be made. (Keene and Holmes, 1987)

Often the customized test is designed to be used for both criterion and norm-referenced purposes. These varying types of information are needed at different levels. State legislatures and state education agencies often require norm-referenced scores for district and schools for monitoring purposes, while objective or criterion referenced information may be more useful to teachers and their students. Test users want to measure their own objectives to increase specific instructional information, as well as, make national comparisons. When tests are used for multiple purposes the information they can provide is often not ideal.

The usefulness of information from the customized test depends on the type of information reported. Keene and Holmes (1987) report that when objective-referenced information is not ideal it tends to be evident to the user. If not enough items are used to assess an objective or an objective is not assessed, these problems are usually taken into account by the interpretation. Users are usually more familiar with assessment of objectives and are often participants in the objective-referenced item development process. This is not often the case with norm-referenced interpretation and development. Users are not involved in the development of the items or in the scales which make interpretation possible. When the information provided for norm-referenced interpretation is less than ideal then the potential for misuse of that data is large.

Interpretation of Customized Test Results -

For the sake of efficiency, customized test results often are used to serve too many purposes, hence the use of results is not ideal for both criterion and normative purposes. Misinterpretation most commonly occurs with the interpretation of norm-referenced results of customized tests. Although the normative information may not be ideal and misinterpretation is possible, the benefits of test customization often outweigh the drawbacks. Users and developers of customized tests need to keep a few factors in mind when making norm-referenced interpretations.

1. In Vitro vs. In Vivo Analysis

In the Biological Sciences a distinction is made of whether a study is carried out "in vitro" (in a laboratory) or "in vivo" (in a living organism or setting). It is important to understand that test equating methods were developed in vitro, but are applied to a dynamic or in vivo situation. Because tests can be equated mathematically does not mean that the interpretation of scores can be the same between a norm-referenced test and its customized derivative. In fact, classical test theory maintains very stringent requirements for equated tests and test scores. In almost all instances, customized tests do not meet these requirements. We are faced with bending the rules, if we are to take advantage of customized testing benefits.

One reason we have very little knowledge about how far the equating rules can bend is that school achievement testing is not easily subjected to laboratory conditions. All research must be pursued "in vivo." Controlling conditions is not feasible at this level. It is very difficult to subject school children to experimental treatments on a large scale. The design used to equate one test to another in a statewide program has to fit into the statewide assessment. Often balanced test administrations and ideal time of testing in terms of the empirical norm dates are difficult constraints to meet. Since conditions are not often controlled and experimental groups are not usually possible, learning from these large scale state and district assessments is often difficult.

2. Content Differences

A large part of the lack of clarity about the usefulness of these estimated norms has to do with the content match issue. Item response theory has provided a method to scale items so that many different combinations can be used, but the level of item content similarity that is necessary in order to scale items is not clear, nor is it evident how the content match affects norm-referenced scores.

Research addressing these questions has not been conclusive. Many of the recent studies examining the match between a set of local objectives and the content of a nationally standardized achievement test suggest that the degree of match can have a fairly large impact on norm-referenced test scores (Bianchini, 1978; Cooley & Leinhardt, 1980; Rentz, 1982). However, the results reported by Phillips and Mehrens (1986) suggest the opposite. They found that the match between local curricula and what is assessed by a nationally norm-referenced test did not seem to have an impact on student performance on the norm-referenced test. Although this research is useful in determining the effects of differing curricula on norm-referenced scores, the research does not address the specific issue of providing norm-referenced test information from tests designed for objective-referenced interpretation.

Examination of the dimensionality of test content may provide some useful information about content differences. A hierarchy of objectives or test content would be useful in establishing expected content dimensions. If the content of the norm-referenced test could be classified then only the customized test content that matched could be used in producing norm-referenced scores.

It is possible to look at test item data and determine more than one dimension. The issue is whether these distinctions between test item content can be made exact enough to fit content examination expectations. If the dimensionality of achievement test content can be explained, test equating may be possible according to the restrictions cited in the literature. Only parallel content would be equated. Norm-referenced scores would only be produced on the basis of a common dimensional structure between the customized test and its base test.

3. Individual vs. Group Interpretation

Individual derived norms on customized tests tend to be less reliable than group scores on the same derived norm scales. When individual scores on norm-referenced scores are compared to individual scores on their customized derivatives the differences are usually higher than the SEMs for the tests, but not much higher. Group scores tend to be relatively stable, if compared to individual distributions. Since group interpretations are often made by comparing means to individual rather than group norms, the risk of misinterpretation is much less when norms are provided only by groups.

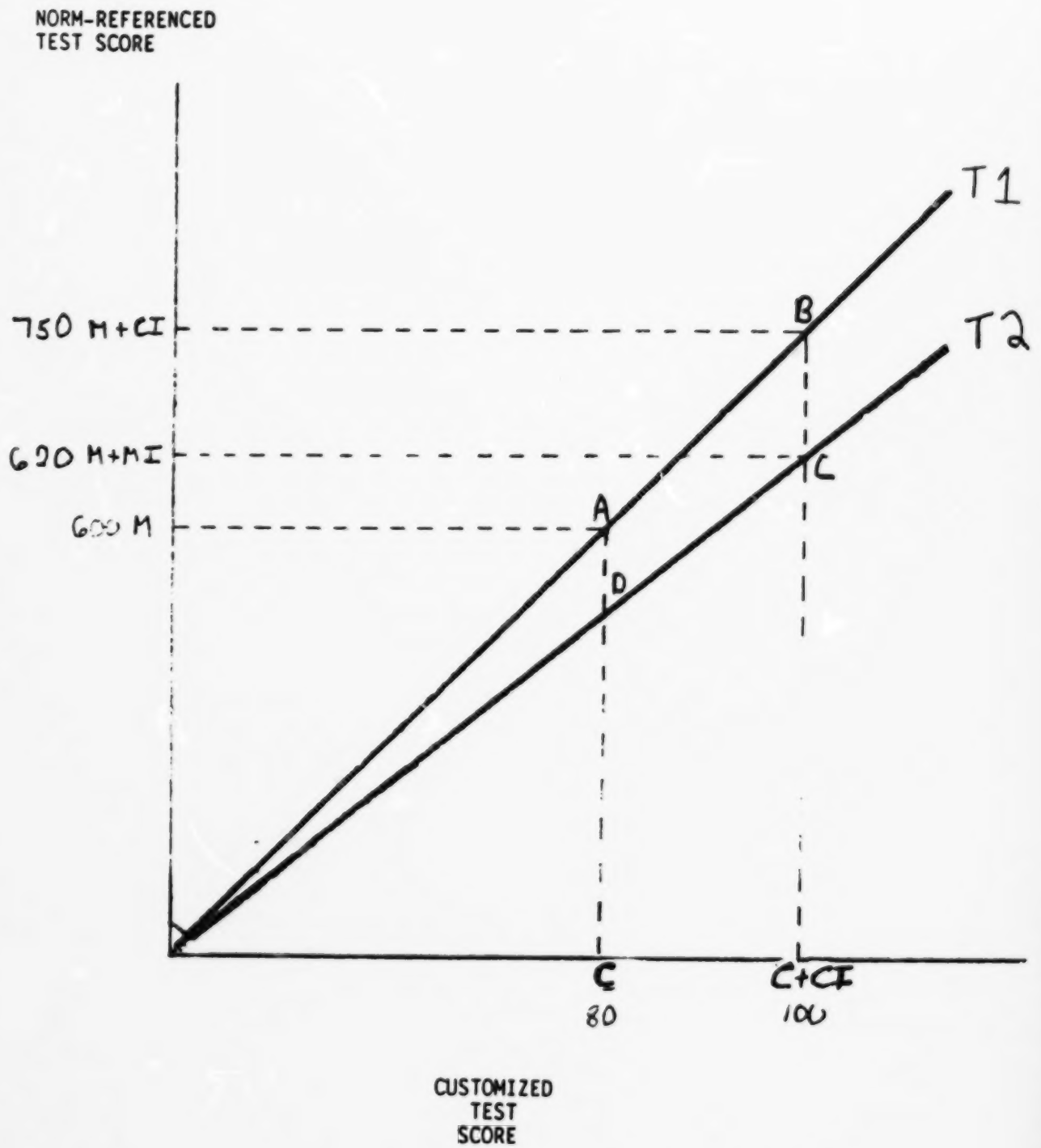
The issue of providing norm-referenced scores for customized tests is, however, not just a score reliability problem. The validity of the scores are also in question. The content differences between norm-referenced and customized assessments determine how valid the interpretation of the scores will be.

Problems with the validity of score interpretation apply equally to groups and individuals, while the reliability issue is greater for individuals than for groups. This reliability difference, which makes group interpretations more reliable, is an artifact of our practice of using individual norm distributions for group interpretation.

4. Differential Instructional Effects Over Time

Providing norm-referenced scores for customized tests poses problems in score interpretation when the equating of the two tests is part of the assessment design. Even more caveats for interpretation are needed when the equating takes place at a different time than the customized test administration. One of the purposes of customization is to focus on instructional goals, and hopefully these attempts will result in better student performance. Since customized test differ in content emphasis from norm-referenced tests because of specific instructional goals, we would expect differential performance increases over time. These increases should favor the customized test. Over time the customized test becomes easier relative to the norm-referenced instrument. This results in inflated norm-referenced scores. Figure 1 illustrates this problem.

FIGURE 1
EQUATING LINE



Conclusion

A growing need for more specific assessment of educational goals has led to the adaptation of norm-referenced assessment tests in order to meet those goals. These adaptations have been termed customized tests. The validity and reliability of score interpretation has been questioned when customized tests have had to serve multiple purposes. Since customized tests provide many benefits to the user, the practice cannot be abandoned, however, each case of customization has to be examined and particular problems for score interpretation need to be identified. Providing norm-referenced score interpretations for customized tests can be allowed if a proper examination of the possible misinterpretations are made.

Text References

- Airasian, Peter W. and Madaus, George F., "Linking Testing and Instruction: Policy Issues." Journal of Educational Measurement, Volume 20, No. 2, Summer 1983, pp 103-118.
- Bianchini, J.C. Achievement tests and differential norms. In M.J. Wargo & D.R. Greene (Eds.), Achievement testing of disadvantaged and minority students for education program evaluation. Monterey, CA: CTB/McGraw-Hill, 1978.
- Cooley, W.W., & Leinhardt, G. The instructional dimensions study. Educational Evaluation and Policy Analysis, 1980, 2, 7-25.
- Haertel, Edward and Calfee, Robert. "School Achievement: Thinking About What To Test." Journal of Educational Measurement, Volume 20, No. 2, Summer 1983, pp 119-132.
- Keene, John M. and Holmes, Susan E., "Models for Obtaining Norm-Referenced Test Information for Statewide (Local) Objective-Referenced Tests: Issues and Challenges." A paper presented at the Annual Meeting of the National Council on Measurement in Education in Washington, D.C., April 22, 1987.
- Linn, Robert L., "Testing and Instruction: Links and Distinctions." Journal of Educational Measurement, Volume 20, No. 2, Summer 1983, pp 179-189.
- Mehrens, W.A. National tests and local curriculum: Match or mismatch? Educational Measurement: Issues and Practices, 1984, 3(3), 9-15.
- Millman, Jason and Arter, Judith A., "Issues in Item Banking." Journal of Educational Measurement, Volume 21, No. 4, Winter 1984, pp 315-330.

Phillips, S.E., & Mehrens, W.A. Relating achievement test scores and item statistics to instructional validity. A paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C., April 1987.

Rentz, R.R. Using the Rasch model to develop curriculum-referenced norms. A paper presented at the Annual Meeting of the National Council on Measurement in Education, New York City, March 1982.

Snow R.E., "Aptitude and Achievement." In W.B. Schrader (Ed.), Measuring Achievement: Progress Over a Decade. New Directions for Testing and Measurement (No. 5). San Francisco: Jossey-Bass, 1980. (Proceedings of the 1979 ETS Invitational Conference.)